[Published in Proceedings of the First International Symposium on *Towards a Comprehensive Intelligence Test (TCIT), Reconsidering the Turing Test for the 21st Century*, AISB 2010 Convention, pp. 7-11, 2010. ISBN: 1902956974]

Causal and Communal Factors in a Comprehensive Test of Intelligence

Paul Schweizer[1]

Abstract. The paper traces a pathway through the existing space of argumentation surrounding the original Turing Test (TT) and the discipline of 'Strong' Artificial Intelligence that followed on from Turing's work, and extends this path to motivate a strengthened conclusion regarding what would be required by a truly Comprehensive Intelligence Test (CIT). The paper begins by examining the initial 'intelligence test' as proposed by Turing, and Searle's high profile critique of both the TT and Strong AI. In tracing the ensuing dialectic between Searle and his own critics, I support Searle's rejection of the 'Systems Reply', and for reasons based more on the philosophical views of Putnam agree that the original Turing Test is fundamentally inadequate. The situation becomes more complex with the 'Robot Reply' and allied Total Turing Test (TTT), and I argue that Searle's attempted refutation of the combined Robotic-Systems reply is unconvincing. However, this is not to say that the position expressed by Searle's opponents is itself confirmed, and I argue that externalist views in the philosophy of language first put forward by Kripke and Putnam cast serious doubt on the issue.

In turn, the causal and communal factors highlighted by externalist views in the philosophy of language point to the need for a fundamental shift in conceptual perspective and a strengthening of criteria in a truly Comprehensive Intelligence Test. I argue that an ideal CIT should focus on the *category* of cognitive system as a whole, rather than on the performance of individual artefacts. From this expanded perspective, the central question is not whether an isolated agent could simulate human performance within the context of a pre-existing sociolinguistic culture developed by the *human* cognitive type. Instead the key issue is whether the artificial cognitive type *itself* is capable of producing a comparable sociolinguistic medium of intelligence, where this essential medium is simply taken for granted as a precondition of the individual performances evaluated in the TT and TTT.

1 THE TURING TEST AND 'STRONG' AI

What would be required for a computational artefact to count as genuinely intelligent in manner comparable to a human being? In 1950 Alan Turing [1] famously proposed an answer to this question, and the controversy launched by his position is still underway. Turing replaced his opening question 'Can (or could) a machine think?' with the more precise and empirically tractable question 'Can (or could) a machine pass a certain type of test?', where the test criteria are framed in terms of *behaviour* that is standardly held to signify intelligence in the case of human beings. In particular, the original 'Turing Test' (TT) is based entirely on *linguistic* inputs and outputs, and is designed as a free ranging session of questions followed by anonymous verbal responses. Linguistic performance is an apt choice as a pivotal criterion of intelligence, since human language is perhaps our most distinctive feature as cognitive agents, and it is an essential medium through which most of our higher level mental achievements are developed and expressed. Hence human language will retain a central role throughout the ensuing discussion.

The TT itself is a disputed issue and many of Turing's successors in the field of Artificial Intelligence would independently endorse a basic

[1] School of Informatics, Univ. of Edinburgh, EH8 9AD, UK. Email: paul@inf.ed.ac.uk.

computational theory of the mind, wherein mentality is held to be explained via the physical realization of the right type of abstract computational procedure, without accepting the TT as an adequate criterion for deeming that a machine possesses genuine intelligence. The two issues are clearly separable, and one can embrace a computational approach to the mind without accepting Turing's original and quite controversial standard. According to the wider project embraced by 'Strong' AI and the computational paradigm, the relation between the abstract program level and its realization in physical hardware could still give an answer to the problematic nature of the relation between mind and body: the mind is to the brain as a program is to the hardware of a digital computer. On this model, computation could then be seen to provide the scientific key for explaining mentality and intelligence. Cognition in general (including the human case) is to be literally described and understood in computational terms.

2 SEARLE'S BASIC CRITIQUE

Probably the most high profile criticism, both of the TT in particular and the computational paradigm in general, was provided by the philosopher John Searle in his 1980 paper 'Minds, Brains and Programs' [2], where Searle put forward his celebrated Chinese Room Argument (CRA), designed to refute the view that passing the TT is a sufficient condition for genuine understanding or intelligence, as well as the more wide-ranging position that a system can sustain genuine mental states in virtue of instantiating the right type of abstract computational procedure. According to Searle, computation per se is neither a necessary nor a sufficient condition for the presence of mental states and real intelligence, and so he deems computation to be fundamentally irrelevant to the issue. Hence both the original TT and the more general mind/program analogy are summarily rejected.

As is well known, the CRA is based on a thought experiment in which Searle, a native speaker of English who knows no Chinese, is locked in a room with a massive rule-book written in English. He receives Chinese inputs on bits of paper and mechanically follows the instruction manual to produce outputs in Chinese script. For the sake of argument, we are asked to suppose that the manual is so good that he is able to fool native speakers of Chinese. They are outside the room giving him the inputs which are questions in Chinese, and by following the recipes for symbol manipulation provided in his rule book, Searle is able to produce appropriate Chinese responses, and on the basis of these outputs, his questioners conclude that a fluent Chinese speaker is locked inside the room.

But Searle doesn't understand Chinese, and doesn't even know basic Chinese vocabulary. He's just mechanically manipulating random 'squiggles and squoggles' according to a program of rules, while the inputs and outputs are, to him, totally meaningless. He has no idea what he's been asked or what he's 'answered'. As far as he's concerned, it's just unintepreted syntax with no intentional content. For example, suppose one of the Chinese questions asked him what was his favorite food, and the output he gave was the Chinese word for 'hamburger'. But he has no idea what the word means, and he comprehends nothing of Chinese, even though he's just passed the Chinese Turing test by serving as a human implementation of the conjectured program. And this is radically different than the case of English, a language Searle really does understand. He *knows* what a hamburger is – a particular kind of greasy foodstuff easily obtained at MacDonalds and Burger King. And this knowledge is a specific property of Searle's mind, a genuine intentional state of understanding the word, and it's not equivalent to any pattern of mere verbal input/output behavior.

Searle's broad sweeping conclusion: mere success at the 'imitation game' and passing the TT is theoretically inadequate as a criterion of intelligence, and instantiating a computer

program has nothing to do with genuine understanding or mentality. But in response to Turing's original question 'can a machine think?' Searle says the answer is definitely y*es* – because *we* are biological machines, and we can think. According to him, we can think in virtue of the unique physical structure and real causal powers of the actual human brain – not in virtue of some abstract *formal shadow*, either classical or connectionist.

3 THE SYSTEMS REPLY

In the style of Turing's 1950 article, Searle goes on to consider and dismiss a number of objections to his view, the first of which he dubs the 'Systems Reply'. A defender of the computational paradigm might argue that perhaps Searle in isolation doesn't understand Chinese, but that's not the point, because the *whole system* that produces the behavior – room plus manual plus Searle – does understand Chinese. Searle responds by claiming that *he* is the only locus of understanding in the scenario, and if he doesn't understand Chinese, then nothing else about the system does. It's absurd to imagine that even though he himself doesn't understand Chinese, somehow the conjunction of Searle plus pencil, slips of paper, instruction manual, four walls, etc. understands Chinese. His pencil and the four walls don't understand *anything at all*. Furthermore, suppose that Searle were gifted with a photographic memory, and could memorize the rule book. Then the entire set-up could be internalized, and Searle could perform the rule governed manipulations simply by consulting his memory, sitting outside under a tree. Searle himself would then *be* the whole system but he still wouldn't understand Chinese.

At this stage I would agree with Searle that the System's reply is not an adequate rejoinder for defending the original TT, mainly because (i) the rest of the system in the original CRA scenario isn't *doing* the right sort of thing and (ii) the standard TT is woefully inadequate in any case (but not because of the introspective angle proffered by Searle – more on this later). At this juncture I would concur with Hilary Putnam's [3] basic critique of the TT: passing the test is not an adequate criterion for concluding that the computer genuinely refers to anything with the strings of symbols it produces, because the computer doesn't have the right sort of relations and interactions with the objects and states of affairs *in the real world* that its words are supposed to be about. If the computer has no eyes, no hands, no mouth, and has never seen or eaten anything, then it is not talking about hamburgers when its program generates the string of English symbols 'h-a-m-b-u-r-g-e-r-s' – it's merely operating inside a closed loop of syntax.

In sharp contrast, *our* talk of hamburgers is intimately connected to *nonverbal* transactions with the objects of reference. There are 'language entry rules' taking us from nonverbal stimuli to appropriate linguistic behaviours. For example, when given the visual stimulus of being presented with a pizza, a taco and a kebab, we can produce the salient utterance 'Those particular foodstuffs are not hamburgers'. And there are 'language exit rules' taking us from linguistic expressions to appropriate nonverbal actions. We can follow complex verbal instructions and produce the indicated patterns of behaviour, e.g. finding the nearest Burger King on the basis of a description of its location in spoken English. Mastery of both of these types of rules is essential for deeming that a human agent understands natural language and is using linguistic expressions in a correct and referential manner - and the hapless TT computer lacks both.

4 THE TOTAL TURING TEST

Hence the standard TT is fundamentally inadequate as a test for understanding, because the range of behaviour it takes into account is far too limited. It relies solely on *verbal* input/output patterns, and these alone are insufficient to ground an interpretation of the manipulated strings. Language is primarily about *extra-*

linguistic entities and states of affairs, and there is nothing in a cleverly designed program for pure syntax manipulation which allows it to break free of this closed loop of symbols and establish a correlation between word and object. When it comes to judging human language users in normal contexts, we rely on a far richer domain of evidence.

So, this criticism suggests a vital strengthening of the TT, later dubbed the Total Turing Test (TTT) by Stevan Harnad [4], wherein the repertoire of relevant behavior is expanded to include the full range of intelligent human activities. This will require that the computational procedures respond to and control not simply a teletype system for written inputs and outputs, but rather a well crafted artificial body. Thus in the TTT the scrutinized artefact is a *robot*, and the data to be tested coincide with the full spectrum of behaviors of which human beings are normally capable. In order to succeed, the test candidate must be able to do, in the real world of objects and people, everything that intelligent people can do. This combined linguistic/robotic test obviously constitutes a vast improvement over Turing's original version, and the range of empirical evidence now encompasses all those forms of complex and varied interaction applicable in the case of our fellow humans. Is the passing the much more rigorous TTT a sufficient condition for deeming that the artefact is truly intelligent? Harnad and others certainly think so, but unsurprisingly, not everyone agrees.

As it happens, the TTT is already anticipated by Searle in his 1980 article, and in his 'Robot Reply' to the CRA he dismisses even this elevated standard with the following line of argument. Imagine that the program doesn't just control verbal responses to verbal inputs as in the TT. Instead, it controls a robot with an artificial body that can successfully behave in the real world just like a person. Surely a device that can pass this extended TTT should count as having a mind? Searle's response: the addition of perceptual and motor capabilities adds nothing to the issue of genuine understanding, *if* these capabilities are controlled by symbolic processes. Accordingly, we can augment the negative CRA thought experiment and suppose that Searle is locked in the room, and now some of the Chinese characters he receives are codes for digitalized inputs from the robot's sensory transducers, and some of the output symbols now control the motors inside the robot's body and make it move its arms and legs. Still, all Searle is doing (perhaps remotely, from inside a control room), is manipulating uninterpreted syntax. He has no idea what is going on outside the control room and the manipulated syntax has no intentional content. Searle cannot *see* the tempting hamburger that the robot's photographic sensing apparatus has transduced into Chinese code, nor is he *trying to grasp it* by outputting the salient effector code controlling the robotic hand.

At this juncture I would take Searle's response to be a plausible answer to the question of whether or not *he* personally understands Chinese, but it is now far from clear that this is the relevant issue. Unlike the case of the standard TT, many of the pivotal inputs and outputs in this more demanding case are no longer manipulated directly by Searle. In order to pass this much more stringent test, the artefact, when viewed as a *system*, must perform physical behaviors in accordance with the language entry and language exit rules appropriate to a genuine understanding of Chinese. Hence the input/output boundaries for the system extend crucially beyond Searle the homunculus. The robot's sensing devices will comprise relevant input boundaries, while its artificial body and limbs will constitute the salient output interface for manifesting the scrutinized behaviour.

So if we apply the systems approach to the robot that passes the TTT, then the situation is no longer comparable to the original TT/CRA scenario, wherein Seale had direct contact with the inputs and outputs under evaluation. Yet Searle attempts to re-employ the same polemical strategy as before, by making the somewhat dubious claim that he could in principle realize the entire system himself and still not understand

Chinese [5]. I am not convinced that this claim expresses an authentic theoretical possibility, but even if we grant Searle's hypothesis for the sake of argument, I would hold that it's still not sufficient to establish his overall conclusion. If Searle could conceivably realize the system and *become* the robot following its instructions, then perhaps Searle would not have introspective access to himself as the realization of a system that understands Chinese – one's intuitions become fairly stretched at this point. But still, this conjectured lack of introspective access does not imply that the *system* does not in fact understand Chinese. It only shows that Searle would not be aware of this fact, which is not a decisive allegation, since clearly there are many aspects of Searle the highly complex real system of which he is personally unaware. And unlike the case of his Systems reply to the original TT, now the system *is* doing exactly the right sort of things, and the test itself is no longer passable while remaining insulated within a mere syntactic bubble.

Hence I would conclude that Searle's introspective considerations are not adequate for deflecting the combined systems-robotic reply. The conjectured fact that Searle might somehow realize the entire system and become the robot following its program, while still lacking the subjective awareness of Chinese semantics, does not establish that the robotic system doesn't understand Chinese. But this is only to say that the argument offered by Searle is unsuccessful at refuting his opponent's claim that the robot possess genuine intelligence and understanding – it is not to say that the positive claim itself has thereby been substantiated. And indeed, I think that other considerations still tend to seriously undermine the view that the successful TTT robot understands language in a manner at all comparable with the paradigmatic human case, and that the expressions generated by the computational artefact are genuinely referential. In contrast to the simplistic TT scenario, the robot can now exhibit mastery of the appropriate language entry and language exit rules, and these are clearly a necessary condition for the referential use of language. But are they sufficient?

5 SEMANTIC EXTERNALISM

Externalist views in the theory of meaning and reference originally put forward by Saul Kripke [6] and Hilary Putnam [7], and subsequently elaborated by Tyler Burger [8], would seem to highlight essential features of natural language (NL) semantics not present in the case of the TTT artefact as currently depicted. The conclusion of Putnam's influential Twin Earth Argument (TEA) is that the internal cognitive states of individual language users radically underdetermine linguistic meaning – there's nothing in the head strong enough to fix reference for terms in natural language.

On Putnam's account, the traditional 'psychologistic' approach ignores two essential aspects of meaning and reference. One (1) is the role of direct causal interaction with the environment when language is acquired and used: natural kind terms such as 'water', 'aluminum', 'gold', 'tiger', etc., make indexical appeal to actual specimens or paradigm cases *in the world* – so causal relations via perception, demonstrative pointing and utterance production in the intersubjectively accessible public domain determine what these words actually refer to. There is no internal encoding or representational state sustained by the individual agent that is powerful enough to do this. According to Putnam's externalist account, it is a semantical fact, quite independent of the mental states of any individual speakers, that 'water' in English means 'the liquid with the same underlying physical microstructure as the stuff in our environment that we interact with when we use the word "water"'. Accordingly the word referred to *this* particular liquid even before we knew that the relevant molecular structure is actually H_2O.

Second (2), the traditional internalist approach ignores what Putnam calls the 'division of linguistic labor': the reliance on *experts* who

set the standards for the entire linguistic community and underwrite the reference relation for natural language in cases where relevant microstructures and/or objective membership conditions *are* known. It is by *acquiring* a natural language within a particular sociolinguistic community and using it within this shared framework that we are able to refer successfully. For example, the average English speaker can use the word 'gold' to talk about real gold, even though they may not know the periodic table, may not know that gold is the element with atomic number 29, and probably don't not know in practice how to distinguish real gold from chalcopyrite. Most people have had causal/perceptual interactions with samples of the metal itself, and thereby have direct indexical access to the substance the word names. But the precise and technical details of the extension of the word 'gold' are uncovered by relevant experts in the field, and it is upon their expertise that our linguistic practice implicitly depends, and not upon our own internal representations or concepts.

In short, language is a communal, historically evolved phenomenon, where the meaning of words is not determined by individuals, but is a public, external matter. Putnam concludes that we must give up the view that meanings are concepts or mental entities of any kind. According to his famous slogan 'Slice the pie any way you like, meanings just ain't in the head.'

6 ROBOTIC REFERENCE

But if meanings ain't in the heads of individual human agents, then they're certainly not in the data bases of computational artefacts. So, in light of (1) above, if the robot's natural language capabilities are simply installed as part of its overall program, then it will not have the necessary history of causal interactions with the objects of reference, and its symbolic activities will remain semantically ungrounded. On the foregoing widely accepted model of 'direct' reference, there is an essential chronological link operating in two directions that semantically tethers an individual's linguistic behaviour to its environmental context.

The relation of reference is founded on a history of causal interactions between the agent and the entities and states of affairs in the world that it uses language to talk about and describe. The word 'water' as used by normal human agents is intimately linked to a long history of associations based on experiences of seeing, drinking, washing with, and being immersed in various samples of environmental H_2O, where these experiences are all *caused* by the liquid itself, giving the agent direct indexical access to water, as the word was acquired and integrated into its overall linguistic framework. And in the other direction, the speaker, when learning language and then applying it proficiently, has repeatedly associated the term 'water' with the liquid so accessed, through bodily and allied verbal ostension ('look, there's a pool of *water* over there'), and intentionally uses the word to pick out the liquid with which it has this history of causal/perceptual episodes.

At this moment in time it's obviously rather difficult to envision exactly how a robot might be designed to pass the TTT, but if its ability to speak fluent English is simply implanted via some sophisticated NLP program, then the concomitant lack of an historical chain of interaction with the real world poses a serious theoretical question regarding the semantical import of its linguistic input/output behaviour. When a token of the term 'water' is emitted by the robot, all shiny and fresh off the assembly line, does it genuinely *mean* 'the liquid with the same underlying microstructure as the stuff in our environment that we interact with when we use the word'? If part of its test were to discourse convincingly on the topic of current theories in the philosophy of language, then it would certainly *say* that it did. But that's a different matter.

Of course, since the robot must be able to behave in all the appropriate manners with its artificial body, then after it's been around for awhile it will have acquired a history of direct causal interaction with water, and in order to pass the test, it must behave *as if* it has made all the associations required to ground its symbolic processes. So I do not present the issue of (1) as an insurmountable obstacle or a conclusive in principle objection, but rather as an interesting and potentially important case of dissimilarity with the semantic analysis of naturally occurring cognitive systems.

However, I think that (2) above presents a much more serious difficulty when evaluating the robot, and one which, even if it could possibly be overcome in the case of an individual artefact, nevertheless suggests that this would still not be enough to attain full parity with humans in the general case. Hence consideration of factor (2) will then serve to motivate the further claim that even the combined linguistic/robotic TTT is intrinsically too limited, and that a conceptual shift in goal posts is required for a truly Comprehensive Intelligence Test (CIT). But first factor (2) itself will be explored in more detail.

In line with Putnam's observations regarding NL semantics, for the robot's linguistic activities to be genuinely referential in a manner comparable to a human being, the robot would have to *acquire* its linguistic fluency through interaction not just with its environment, but as a member of the relevant sociolinguistic community. And again, this is very different than having its language processing abilities simply programmed in as a finished product, particularly if this finished product were predesigned in terms of some particular external target language.

If the robot did not l*earn* its language via extended participation with an actual and embodied linguistic culture, within a shared physical and social context, then it will not be a valid member of any such community, and consequently it will be unable to rely upon the division of linguistic labour central to our referential success. Putnam gives the analogy that natural language is not like a hammer, a tool that can be wielded successfully by an individual. Instead, language is a cooperative social venture, more like operating a steam ship or perhaps a large industry. As bone fide members of the English speaking 'linguistic cooperative', we're automatically plugged in to this ancient and highly structured communication system, a living cognitive network through which we inherit and access the meaning of our words.

For the linguistic activities of single human beings to be semantically grounded, the individuals must belong to and participate in such a communication network, a network that is anchored to a continuous presence extended in both time and space. People first have direct causal interactions with various persons, places, objects and natural kinds in their immediate surroundings, and by learning and exercising their linguistic behaviours in this shared environment, they enjoy direct indexical access to the referents of the corresponding terms. But via full membership in this same NL community, they also gain linguistic access to people, places, objects, substances and states of affairs *remote* in both time and space. I've never been to Madagascar, and Isaac Newton died long before I was born. Nonetheless, through membership in the English speaking NL sociolinguistic coop, I'm plugged into this ongoing, far reaching and extremely powerful communication network, and am able to use English words to successfully *talk about* Isaac Newton and Madagascar, even though I've been in direct personal contact with neither.

However, if the TTT robot's English speaking abilities are simply installed as part of some highly sophisticated NLP software package, then it will lack the essential history of having acquired these abilities through interaction and participation in an actual and embodied community. Its 'semantics' will be purely internal and solipsistic, tethered to files stored in its data bases and various coded representations supplied by its designers. And as Putnam's TEA convincingly shows, such internal states and

structures are incapable of determining the reference relation for even such basic natural kind terms as 'water'.

Of course, in the same vein as noted above, the combined linguistic/robotic standards of the TTT would require the robot to have extended dealing with human beings while it was undergoing the test, and one might then argue that after it had been around for some time and had sufficient verbal and other behavioural interchanges with humans, it would itself gradually *become* a card carrying member of the English speaking sociolinguistic coop, with full rights and privileges. And while a case perhaps could be made that a successful TTT robot, fully integrated into human society, might eventually be deemed a legitimate member of the English speaking community, the issue nonetheless points to a fundamental feature of intelligent human behaviour that seems entirely absent in the standard test scenarios considered so far, and which this form of mere *integration* would fail to address.

7 A TRULY COMPRENSIVE TEST

Human intelligence as we know it depends in an essential manner on membership in a linguistic, intellectual community, and furthermore, one that has been created and is sustained by *conspecifics*: the intelligent behavior of human individuals is inseparable from immersion in a historically evolved culture of intelligence, where this culture is itself the product of *human* cognitive processing. Thus human intelligence as an indigenous phenomenon is not merely individualistic, but rather presupposes for its development and expression essential involvement in a specialized social context that is itself a product of the *human* cognitive *type*.

In this sense it is asymmetrical for a truly Comprehensive Test of Intelligence to focus merely on the performance of individual artefacts, rather than on the overall capabilities of the cognitive type to which these individuals belong. So the manner in which the much more rigorous TTT is envisaged still reveals a crucial disanalogy with the human case. Not only can individual human beings exhibit the salient patterns of verbal input/output behaviour required by the original Turing Test, and full blown mastery of the language entry and exit rules required by the combined linguistic and robotic Total Turing Test, but it was the *human* cognitive type, of which such human individuals are members, that has produced natural language and this advanced culture of intelligence in the first place. And it is with other tokens of this *same* type that we intermingle as a sociolinguistic community and upon whom our referential success co-depends.

In sharp contrast, the computational artefact involved in the TTT is *not* a member of this same type. It has an alien and artificial cognitive structure that is quite possibly incapable, at the type level, of ever producing natural language or the kind of sociolinguistic context which is simply *presupposed* as a starting point by the TTT. It is given a prefabricated stage on which to perform acts of post hoc imitation, and this kind of test could conceivably be passed by a well designed puppet, rather than a robust and genuinely intelligent category of cognitive system.

On these grounds the TTT is still too weak, because an individual artefact merely has to perform successfully in a pre-existing natural language community, in a context and medium of intelligence produced by a radically different cognitive type - the same type as its designers! Its behavioural outputs can presuppose sophisticated, pre-structured linguistic inputs for free, and these can serve as triggers for appropriately complex responses. And these factors make it ambiguous as to whether the locus of genuine intelligence resides in the designers or in their artefact. However, *human* cognitive architecture was first tested on primitive and pre-linguistic environmental inputs that were transformed by members of this type over tens of thousands of years to yield the sophisticated sociolinguistic community presupposed by the

robot (see the discussion in my [9] for allied points motivated by a somewhat different set of considerations)

8 CONCLUSION

A truly Comprehensive Intelligence Test (CIT) would require the artefact's cognitive architecture to start from scratch, with the same primitive inputs as our pre-linguistic forebears. And this is why the standard science fiction scenario of an advanced alien life form, regardless of its chemical composition or internal processing structure, is always a more convincing hypothetical case of true intelligence than a puppet-like TTT artefact. In contrast to a robot, the alien life form must have evolved its own sociolinguistic culture of native intelligence, in response to its primitive environmental stimuli, rather than exhibiting programmed capabilities *simulating* real intelligence in a pre-existing context for which it was tailor made by its designers. In the speculative case of an advanced alien life form, the *type* of cognitive architecture in question would already have passed a CIT.

So if individual tokens of this alien type become fluent in English and are then able to interact successfully with humans and pass an interplanetary version of the TTT, then we would clearly be warranted in concluding that the individual specimens in question passed the TTT in virtue of possessing *genuine* intelligence on a par with human beings. The fact that the alien's cognitive type has already passed its own CIT is a necessary background condition for the deployment of the much narrower TTT in the case of particular tokens performing successfully in the context of a sociolinguistic medium created by humans.

So this points to an important shift in conceptual perspective: a truly comprehensive test should focus on the capacities of the category of cognitive system as whole, rather than on the performance of isolated tokens. The original TT was deliberately posed as an *imitation game*, and this is an intrinsic limitation on its adequacy. An individual artefact could in principle pass the TT by simulating verbal intelligence within an extremely sophisticated context already developed by a radically distinct cognitive type. Although a vast improvement in many ways, the TTT still incorporates this intrinsic limitation, by again focussing on a token artefact, specifically designed to mimic the full range of intelligent behaviour within a cultural network produced by a different category of cognitive agent altogether. But rather than consider the imitation of *our* intelligent behaviour by specially designed tokens, the criterion for a CIT should be whether the artificial type itself is capable of producing a comparable medium of intelligence, starting from the primitive environmental inputs of our pre-linguistic forebears. This advanced and essential medium is simply taken for granted as a precondition of both the TT and TTT, yet if an artificially devised cognitive architecture were able to develop such a sociolinguistic culture on its own, this would require and constitute a genuine *manifestation* of intelligence, and would not be an act of mere simulation.

REFERENCES

[1] A. Turing, 'Computing Machinery and Intelligence', *Mind* 59: 433-460 (1950).

[2] J. Searle, 'Minds, Brains and Programs', *Behavioral and Brain Sciences* 3: 417-424, (1980).

[3] H. Putnam, 'Brains in a Vat', in *Reason, Truth and History*, Cambridge University Press, (1981).

[4] S. Harnad 'Other Bodies, Other Minds: A Machine Incarnation of an Old Philosophical Problem', *Minds and Machines* 1: 43-54, (1991).

[5] J. Preston and M. Bishop *Views into the Chinese Room*, Oxford University Press, (2002).

[6] S. Kripke, *Naming and Necessity*, Harvard University Press, (1972).

[7] H. Putnam , 'The Meaning of 'Meaning'', in

Mind, Language and Reality, Cambridge University Press, (1975).
[8] T. Burge, 'Individualism and the Mental', in French, P., Euhling, T., and Wettstein, H. (eds.), *Studies in Epistemology*, vol. 4, *Midwest Studies in Philosophy*, University of Minnesota Press, (1979).
[9] P. Schweizer 'The Truly Total Turing Test' *Minds and Machines* **8:** 263-272, (1998).